

8

Common Knowledge and Nash Equilibrium

Where every man is Enemy to every man... the life of man is solitary, poore, nasty, brutish, and short.

Thomas Hobbes

In the case of any person whose judgment is really deserving of confidence, how has it become so? Because he has kept his mind open to criticism of his opinions ...

John Stuart Mill

This chapter applies the modal logic of knowledge developed in §4.1 and §5.10 to explore sufficient conditions for Nash equilibrium in two-player games (§8.1). We then expand the modal logic of knowledge to multiple agents and prove a remarkable theorem, due to Aumann (1976) that asserts that an event that is self-evident for each member of a group is common knowledge (§8.3).

This theorem is surprising because it appears to prove that individuals know the content of the minds of others with no explicit epistemological assumptions. We show in §8.4 that this theorem is the result of implicit epistemological assumptions involved in the construction of the standard semantic model of common knowledge, and when more plausible assumptions are employed, the theorem is no longer true.

Aumann's famous *agreement theorem* is the subject of section §8.7, where we show that the Aumann and Brandenburger (1995) conditions for Nash equilibrium in multi-player games is essentially an agreement theorem. Because there is no principle of Bayesian rationality that gives us the commonality of beliefs on which agreement depends, our analysis entails the demise of methodological individualism, a theme explored in section §8.8.

8.1 Conditions for Nash Equilibrium in Two-Player Games

Suppose that rational agents know one another's conjectures (§4.1), so that if $\phi_i^\omega(s_{-i}) > 0$ and $s_j \in S_j$ is player j 's pure strategy in s_{-i} , then s_j is a best response to his conjecture ϕ_j^ω . We then have a genuine "equilibrium in

conjectures,” as now no agent has an incentive to change his pure strategy choice s_i , given the conjectures of the other players. We also have

THEOREM 8.1 *Let \mathcal{G} be an epistemic game with Bayesian rational players, and suppose in state ω each player i knows the others’ actions $s_{-i}(\omega)$. Then $s(\omega)$ is a Nash equilibrium.*

PROOF: To prove this theorem, which is due to Aumann and Brandenburger (1995), note that for each i , i knows the other players’ actions at ω , so $\phi_i^\omega(s_{-i}) = 1$, which implies $s_{-i}(\omega) = s_{-i}$ by K3, and i ’s Bayesian rationality at ω then implies $s_i(\omega)$ is a best response to s_{-i} . ■

We say a Nash equilibrium in conjectures $(\phi_1^\omega, \dots, \phi_n^\omega)$ occurs at ω if for each player i , $s_i(\omega)$ is a best response to ϕ_i^ω , and for each i , $\phi_i^\omega \in \Delta^*S_{-i}$. We then have

THEOREM 8.2 *Suppose \mathcal{G} is a two-player game, and at $\omega \in \Omega$, for $i = 1, 2, j \neq i$,*

1. *Each player knows the other is rational: i.e., $\forall \omega' \in \mathbf{P}_i \omega, s_j(\omega')$ is a best response to $\phi_j^{\omega'}$;*
2. *Each player knows the other’s beliefs; i.e., $\mathbf{P}_i \omega \subseteq \{\omega' \in \Omega \mid \phi_j^{\omega'} = \phi_j^\omega\}$.*

Then, the mixed strategy profile $(\sigma_1, \sigma_2) = (\phi_2^\omega, \phi_1^\omega)$ is a Nash equilibrium in conjectures.

PROOF: To prove the theorem, which is due to Aumann and Brandenburger (1995) and Osborne and Rubinstein (1994), suppose s_1 has positive weight in $\sigma_1 = \phi_2^\omega$. Because $\phi_2^\omega(s_1) > 0$, there is some ω' such that $\omega' \in \mathbf{P}_2 \omega$ and $s_1(\omega') = s_1$. By (1) s_1 is a best reply to $\phi_1^{\omega'}$, which is equal to ϕ_1^ω by (2). Thus s_1 is a best reply to $\sigma_2 = \phi_1^\omega$, and a parallel argument shows that s_2 is a best reply to σ_1 , so (σ_1, σ_2) is a Nash equilibrium. ■

8.2 A Three-player Counterexample

	L	R		L	R
U	2,3,0	2,0,0	U	0,0,0	0,2,0
D	0,3,0	0,0,0	D	3,0,0	3,2,0
	W			E	

Figure 8.1. Alice, Bob and Carole

Unfortunately, Theorem 8.2 does not extend to three or more players. For example Figure 8.1 shows a game where Alice chooses the row (U, D), Bob chooses the column (L, R), and Carole chooses the matrix (E, W) (the example is due to Osborne and Rubinstein, 1994:79). Note that every strategy of Carole's is a best response, because her payoff is identically zero. We assume there are seven states, so $\Omega = \{\omega_1, \dots, \omega_7\}$, as depicted in Figure 8.2. States ω_1 and ω_7 represent Nash equilibria. There are also two sets of mixed strategy Nash equilibria. In the first, Alice plays D , Carole plays $2/5W + 3/5E$, and Bob plays anything (Carole's strategy is indeed specified by the condition that it gives Bob equal payoffs for all strategies), while in the second, Bob plays L , Carole plays $3/5W + 2/5E$, and Alice plays anything (this time, Carole's strategy is specified by the condition that it equalizes all Alice's payoffs).

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
P	32/95	16/95	8/95	4/95	2/95	1/95	32/95
s_1	U	D	D	D	D	D	D
s_2	L	L	L	L	L	L	R
s_3	W	E	W	E	W	E	E
\mathcal{P}_A	$\{\omega_1\}$	$\{\omega_2, \omega_3\}$	$\{\omega_4, \omega_5\}$	$\{\omega_6\}$	$\{\omega_7\}$		
\mathcal{P}_B	$\{\omega_1, \omega_2\}$	$\{\omega_3, \omega_4\}$	$\{\omega_5, \omega_6\}$	$\{\omega_7\}$			
\mathcal{P}_C	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_3\}$	$\{\omega_4\}$	$\{\omega_5\}$	$\{\omega_6\}$	$\{\omega_7\}$

Figure 8.2. Information Structure for Alice, Bob, and Carole game. Note that P is the probability of the state, s_i is i 's choice in the corresponding state, and \mathcal{P}_i is the knowledge partition for individual i .

Because there is a common prior (the 'P' row in Figure 8.2), and every state is in the corresponding cell of partition for each player (the last three rows in the figure), these are true knowledge partitions. Moreover, the posterior probabilities for the players are compatible with the knowledge operators for each player. For instance, in state ω_4 , $\mathbf{P}_A\omega_4 = \{\omega_4, \omega_5\}$, and the conditional probability of ω_4 , given $\mathbf{P}_A\omega_4$, is $2/3$, and that of ω_5 is $1/3$. Therefore, Alice's conjecture for Bob is $\phi_{AB}^{\omega_4} = L$, and for Carole is $\phi_{AC}^{\omega_4} = 2/3E + 1/3W$. Alice's move at ω_4 , which is D , is therefore a best response, with payoff 2 as opposed the payoff of $2/3$ from playing U against L and $2/3E + 1/3W$. Moreover, Alice knows that Carole is rational at ω_4 (trivially, because her payoff does not depend on her move). Alice knows Bob's beliefs at ω_4 , because Bob could be either in

\mathcal{P}_B partition cell $\{\omega_3, \omega_4\}$ or $\{\omega_5, \omega_6\}$, in both of which he believes Alice plays D and Carole plays $2/3W + 1/3E$. She also knows that Bob plays L in both cells, and Bob is rational because L pays off 2 against D and $2/3W + 1/3E$, as opposed to payoff $2/3$ to playing R . Similarly, at ω_4 , $\mathbf{P}_B\omega_4 = \{\omega_3, \omega_4\}$, so Bob knows that Alice is in either \mathcal{P}_A partition cell $\{\omega_2, \omega_3\}$ or $\{\omega_4, \omega_5\}$, in both of which Alice knows that Bob plays L and Carole plays $2/3E + 1/3W$. Thus, Bob knows Alice's beliefs and that Alice is rational in playing D . Similar reasoning shows that, Carole knows Alice and Bob's beliefs, and that they are rational at ω_4 . Thus, all the conditions of the previous theorem are satisfied at ω_4 , but of course, the conjectures at ω_4 do not form a Nash equilibrium, because $\phi_{AB}^{\omega_4} = L$ and $\phi_{BA}^{\omega_4} = D$ are not part of any Nash equilibrium of the game.

The reason Theorem 8.2 does not extend to this three player game is that Alice and Bob have different conjectures as to Carole's behavior, which is possible because Carole has more than one best response to Alice and Bob. They both know Carole is rational and they both know Carole believes $\phi_C^\omega = \{D, L\}$, for $\omega \in \{\omega_2, \dots, \omega_5\}$. However, these do not determine Carole's mixed strategy. Thus, mutual knowledge of rationality and beliefs is not sufficient to ensure that a Nash equilibrium will be played.

8.3 The Modal Logic of Common Knowledge

Suppose we have a set of n of agents, each of whom has a knowledge operator \mathbf{K}_i , $i = 1, \dots, n$. We say $E \subseteq \Omega$ is a *public event* if E is self-evident for all $i = 1, \dots, n$. By K1, Ω is a public event, and if E and F are public events, so is $E \cap F$, by K2a. Hence, for any $\omega \in \Omega$, there is a minimal public event $\mathbf{P}_*\omega$ containing ω ; namely the intersection of all public events containing ω .

We can construct $\mathbf{P}_*\omega$ as follows. First, let

$$\mathbf{P}_*^1\omega = \bigcup_{j \in N} \mathbf{P}_j\omega, \quad (8.1)$$

which is the set of states that are possible for at least one agent at ω . Now, ω is possible for all players i from every state $\omega' \in \mathbf{P}_*^1\omega$, but an arbitrary $\omega' \in \mathbf{P}_*^1\omega$ is possible for some player i at ω , although not necessarily for all. So, $\mathbf{P}_*^1\omega$ may not be a public event. Thus we define

$$\mathbf{P}_*^2\omega = \bigcup \{\mathbf{P}_*^1\omega' \mid \omega' \in \mathbf{P}_*^1\omega\}, \quad (8.2)$$

which is the set of states that are possible for some agent at some state in $\mathbf{P}_*^1\omega$; i.e., this is the set of states that are possible for some agent from some state ω' that is possible for some (possibly other) agent at ω . Using similar reasoning, we see that any state in \mathbf{P}_*^1 is possible for any player i and any state $\omega' \in \mathbf{P}_*^2$, but there may be states in $\mathbf{P}_*^2\omega$ that are possible for one or more agents, but not all agents. In general, having defined $\mathbf{P}_*^i\omega$ for $i = 1, \dots, k-1$, we define

$$\mathbf{P}_*^k\omega = \bigcup \{\mathbf{P}_*^1\omega' \mid \omega' \in \mathbf{P}_*^{k-1}\omega\}. \quad (8.3)$$

Finally, we define

$$\mathbf{P}_*\omega = \bigcup_{k=1}^{\infty} \mathbf{P}_*^k\omega. \quad (8.4)$$

This is the set of states ω' such that there is a sequence of states $\omega = \omega_1, \omega_2, \dots, \omega_{k-1}, \omega_k = \omega'$ such that ω_{r+1} is possible for some agent at ω_r , for $r = 0, \dots, k-1$. Of course, this is really a finite union, because Ω is a finite set. Therefore, for some k , $\mathbf{P}_*^k\omega = \mathbf{P}_*^{k+i}\omega$ for all $i \geq 1$.

We can show that $\mathbf{P}_*\omega$ is the minimal public event containing ω . First, $\mathbf{P}_*\omega$ is self-evident for each $i = 1, \dots, n$, because for every $\omega' \in \mathbf{P}_*\omega$, $\omega' \in \mathbf{P}_*^k\omega$ for some integer $k \geq 1$, so $\mathbf{P}_i\omega' \subseteq \mathbf{P}_*^{k+1}\omega \subseteq \mathbf{P}_*\omega$. Hence $\mathbf{P}_*\omega$ is a public event containing ω . Now let E be any public event containing ω . Then, E must contain $\mathbf{P}_i\omega$ for all $i = 1, \dots, n$, so $\mathbf{P}_*^1\omega \subseteq E$. Assume we have proven $\mathbf{P}_*^j\omega \subseteq E$ for $j = 1, \dots, k$. Because $\mathbf{P}_*^k\omega \subseteq E$ and E is a public event, then $\mathbf{P}_*^{k+1}\omega = \mathbf{P}_*^1(\mathbf{P}_*^k\omega) \subseteq E$. Thus $\mathbf{P}_*\omega \subseteq E$.

The concept of a public event can be defined directly in terms of the agents' partitions $\mathcal{P}_1, \dots, \mathcal{P}_n$. We say partition \mathcal{P} is *coarser* than partition \mathcal{Q} if every cell of \mathcal{Q} lies in some cell of \mathcal{P} , and we say \mathcal{P} is *finer* than \mathcal{Q} if \mathcal{Q} is coarser than \mathcal{P} . The public event partition \mathcal{P}_* corresponding to \mathbf{P}_* is then the finest common coarsening of the partitions $\mathcal{P}_1, \dots, \mathcal{P}_n$ of the individual players.

To visualize these concepts, we return to the corn field analogy (§4.1). To coarsen a partition, simply remove one or more fence segments, and then to be tidy, repeatedly remove any fence segments that have either end unconnected to another segment. To refine (i.e., make finer) a partition, simply partition one or more of its cells. If the field has two partitions, visualize one with fence segments colored red and the other with segments colored blue. Where the fence segments intersect, let them share a common fence pole. Where a red and a blue fence segment separate the same corn stalks,

including the fence segments surrounding the whole corn field, merge them into red and blue striped fence segments. The finest common coarsening of the two partitions is then the partition formed by removing all fence segments that are of only one color.

This visualization extends directly to the public event partition corresponding to the knowledge partitions in an n -player game. We give each player's fence partition a distinctive color, and we allow two or more agents to share fence segments by applying multiple colors to shared segments. We allow fence segments of different agents to pass through one another by placing a common fence pole at a point of intersection. Now, remove all fence segments that have fewer than n colors. What remains is the public event partition. Alternatively, the minimal public event $\mathbf{P}_*\omega$ containing state ω consists of the states that can be attained by walking from ω to any state in the field, provided one never climbs over a fence shared by all players.

Clearly the operator \mathbf{P}_* satisfies P1. To show that it also satisfies P2, suppose $\omega' \in \mathbf{P}_*\omega$. Then, by construction, $\mathbf{P}_*\omega' \subseteq \mathbf{P}_*\omega$. To show that $\mathbf{P}_*\omega' = \mathbf{P}_*\omega$, note that $\omega' \in \mathbf{P}_*^k\omega$ for some k . Therefore, by construction, there is a sequence $\omega = \omega_1 = \dots = \omega_k = \omega'$, such that $\omega_{j+1} \in \mathbf{P}_{i_j}\omega_j$ for some $i_j \in n$, for $j = 1, \dots, k-1$. However, reversing the order of the sequence shows that $\omega \in \mathbf{P}_*\omega'$. Therefore $\mathbf{P}_*\omega = \mathbf{P}_*\omega'$. This proves that P2 holds, so \mathbf{P}_* has all the properties of a possibilities operator.

It follows that \mathbf{P}_* is a possibility operator. We define a *public event* operator \mathbf{K}_* as the knowledge operator corresponding to the possibility operator \mathbf{P}_* , so $\mathbf{K}_*E = \{\omega \mid \mathbf{P}_*\omega \subseteq E\}$. We can then define an event E as a *public event* at $\omega \in \Omega$ if $\mathbf{P}_*\omega \subseteq E$. Thus, E is a public event if and only if E is self-evident to all players at each $\omega \in E$. Also, E is a public event if and only if E is the union of minimal public events of the form $\mathbf{P}_*\omega$. Moreover, K5 shows that if E is a public event, then at every $\omega \in E$ everyone knows that E is a public event at ω .

In the standard treatment of common knowledge (Lewis 1969, Aumann 1976), an event is common knowledge if everyone knows E , everyone knows that everyone knows E , and so on. It is easy to see a public event is always common knowledge, and conversely. For, suppose E is a public event. Then, for any $i, j, k = 1, \dots, n$, $\mathbf{K}_iE = E$, $\mathbf{K}_j\mathbf{K}_iE = \mathbf{K}_jE = E$, $\mathbf{K}_k\mathbf{K}_j\mathbf{K}_iE = \mathbf{K}_kE = E$, and so on. Thus all events of the form $\mathbf{K}_k\mathbf{K}_j \dots \mathbf{K}_iE$ are self-evident for k , so E is common knowledge. Conversely, suppose that for any sequence $i, j, \dots, k = 1, \dots, n$,

$\mathbf{K}_i \mathbf{K}_j \dots \mathbf{K}_k E \subseteq E$. Then, for any $\omega \in E$, because $\mathbf{P}_i \omega \subseteq E$, we have $\mathbf{P}_*^1 \omega \subseteq E$, where \mathbf{P}_*^1 is defined in (8.1). We also have $\mathbf{K}_i \mathbf{P}_*^1 \omega \subseteq E$, because $\mathbf{K}_i \mathbf{K}_j E \subseteq E$ for $i, j = 1, \dots, n$, so $\mathbf{P}_*^2 \omega \subseteq E$ from (8.2). From (8.3), we now see that $\mathbf{P}_*^k \omega \subseteq E$ for all k , so $\mathbf{P}_* \omega \subseteq E$. Therefore E is the union of public events, and hence is a public event.

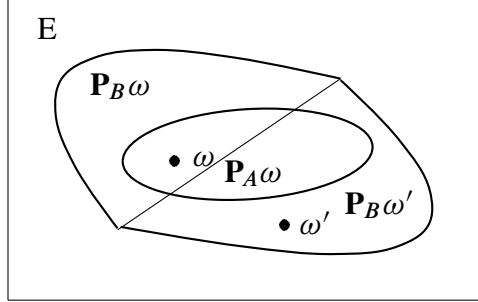


Figure 8.3. At ω , Bob knows that Alice knows that E .

Figure 8.3 shows the situation where Alice knows E at ω , because her minimal self-evident event $\mathbf{P}_A \omega$ at ω lies within E . Moreover $\mathbf{P}_A \omega$ intersects two of Bob's minimal self-evident events, $\mathbf{P}_B \omega$ and $\mathbf{P}_B \omega'$. Because both of $\mathbf{P}_B \omega$ and $\mathbf{P}_B \omega'$ lie within E , Bob knows that Alice knows that E at ω (and at every other state in $\mathbf{P}_A \omega$).

8.4 The Commonality of Knowledge

We have defined a public event as an event that is self-evident to all players. We then showed that an event E is public if and only if, it is common knowledge. It appears, then, that at a public event, there is a perfect *commonality of knowledge*: players know a great deal about what other players know. Where does this knowledge come from? The answer is that we have tacitly assumed that the way each individual partitions Ω is “known” to all, not in the formal sense of a knowledge operator, but rather in the sense that an expression of the form $\mathbf{K}_i \mathbf{K}_j E$ makes sense, and means “ i knows that j knows that E .” Formally, to say that i knows that j knows E at ω means that at every state $\omega' \in \mathbf{P}_j \omega$, $\mathbf{P}_i \omega' \subseteq E$. But, i knows that this is the case only if he “knows” $\mathbf{P}_j \omega$, which allows him to test $\mathbf{K}_i \omega' \subseteq E$ for each $\omega' \in \mathbf{P}_j \omega$.

For example, suppose Alice, Bob, and Carole meet yearly on a certain date at a certain time to play a game \mathcal{G} . Suppose, by chance, all three happened to be in Dallas, Texas the day before, and although they did not

see each other, each witness the same highly unusual event x . We define the universe $\Omega = \{\omega, \omega'\}$, where the unusual even occurs in ω but not in ω' . Then, $\mathbf{P}_A\omega = \mathbf{P}_B\omega = \mathbf{P}_C\omega = \{\omega\}$, and hence $\mathbf{K}_A\omega = \mathbf{K}_B\omega = \mathbf{K}_C\omega = \{\omega\}$. Thus ω is self-evident to all three individuals, and hence ω is a public event. Therefore at ω , Alice knows that Bob knows that Carole knows ω , and so on. But, of course, this is not the case. Indeed, none of the three individuals is aware that the others know the event x .

The problem is that we have misspecified the universe. Suppose an event ω is a four-vector, the first entry of which is either x or $\neg x$ (meaning “not x ”), and the other three entries are “true” or “false,” depending on whether Alice, Bob, and Carole, respectively, knows or does not know whether x occurred. The universe Ω now has sixteen distinct states, and the state ω that actually occurred is $\omega = [x, \text{true}, \text{true}, \text{true}]$. However, now $\mathbf{P}_A\omega = \{\omega' \in \Omega \mid \omega'[1] = x \wedge \omega'[2] = \text{true}\}$. Therefore, the state ω is now *not* self-evident for Alice. Indeed, the smallest self-evident event $\mathbf{P}_A\omega$ for Alice at ω in this case is Ω itself!

This line of reasoning reveals a central lacuna in epistemic game theory: its semantic model of common knowledge assumes too much. Economists have been misled by the elegant theorem that says mutual self-evidence implies common knowledge into believing the axioms of rational choice imply something substantive concerning the commonality of knowledge across agents. They do not. Indeed, there is no formal principle specifying conditions under which distinct individuals will attribute the same truth-value to a proposition p with empirical content (we can assume rational agents will all agree on mathematical and logical tautologies), or will have a mental representation of the fact that others attribute truth-value to p . We address this below by sketching the attributes of what we have termed *mutually accessible* events (§7.8).

8.5 The Tactful Ladies

While walking in the garden, Alice, Bonnie and Carole encountered a violent thunderstorm and are obliged to duck hastily into a restaurant for tea. Carole notices that Alice and Bonnie have dirty foreheads, although each is unaware of this fact. Carole is too tactful to mention this embarrassing situation, which would surely lead them to blush, but she observes that, like herself, each of the two ladies knows that someone has a dirty forehead but is also too tactful to mention this fact. The thought occurs to Carole that she

also might have a dirty forehead, but there are no mirrors or other detection devices handy that might help resolve her uncertainty.

At this point, a little boy walks by the three young ladies' table and exclaims "I see a dirty forehead!" After a few moments of awkward silence, Carole realizes that she has a dirty forehead, and blushes.

How is this feat of logical deduction possible? Certainly it is mutually known among the ladies that at least one of them had a dirty forehead, so the little boy did not inform any of them of this fact. Moreover, each lady could see that the other ladies each saw at least one dirty forehead, so it is mutually known that each lady knew what the little boy said before he said it. However, the little boy's remark does inform each lady that they all know that they all know that one of them has a dirty forehead. This is something that none of the ladies knew before the little boy's announcement. For instance, Alice and Bonnie each knows she might not have a dirty forehead, so Alice knows that Bonnie might believe that Carole sees two clean foreheads, in which case Alice and Bonnie know that Carole might not know that there is at least one dirty forehead. Following the little boy's announcement, however, and assuming the other ladies are logical thinkers (which they must be if they are Bayesian decision-makers), Carole's inference concerning the state of her forehead is unavoidable.

To see why, suppose Carole did not have a dirty forehead. Carole then knows that Alice sees one dirty forehead (Bonnie's), so Alice learns nothing from the little boy's remark. But, Carole knows that Bonnie sees that Carole's forehead is not dirty, so if Bonnie's forehead were not dirty, then Alice would have seen two clean foreheads, and the little boy's remark would imply that Alice would know that she was the unfortunate possessor of a dirty forehead. Because Alice did not blush, Carole knows that Bonnie would conclude that she must have a dirty forehead, and would have blushed. Because Bonnie did no such thing, Carole knows that her assumption that she has a clean forehead is false.

To analyze this problem formally, suppose Ω consists of eight states of the form $\omega = xyz$, where $x, y, z \in \{d, c\}$ are the states of Alice, Bonnie, and Carole, respectively and where d and c stand for "dirty forehead" and "clean forehead," respectively. Thus, for instance $\omega = ccd$ is the state of the world where Carole has a dirty forehead but Alice and Bonnie both have clean foreheads. When Carole sits down to tea, she knows $E_C = \{ddc, ddd\}$, meaning she sees that Alice and Bonnie have dirty foreheads, but her own forehead could be either clean or dirty. Similarly, Alice knows

$E_A = \{cdd, ddd\}$ and Bonnie knows $E_B = \{dcd, ddd\}$. Clearly, no lady knows her own state. What does Bonnie know about Alice's knowledge? Because Bonnie does not know the state of her own forehead, she knows that Alice knows the event "Carole has a dirty forehead," which is $E_{BA} = \{cdd, ddd, ccd, dcd\}$. Similarly, Carole knows that Bonnie knows that Alice knows $E_{CBA} = \{cdd, ddd, ccd, dcd, cdc, ddc, dcc, ccc\} = \Omega$. Assuming Carole has a clean forehead, she knows that Bonnie knows that Alice knows $E'_{CBA} = \{cdc, ddc, dcc, ccc\}$. After the little boy's announcement, Carole would then know that Bonnie knows that Alice knows $E''_{CBA} = \{cdc, ddc, dcc\}$, so if Bonnie did not have a dirty forehead, she would know that Alice knows $E'''_{BA} = \{dcc\}$, so Bonnie would conclude that Alice would blush. Thus, Bonnie's assumption that she has a clean forehead would be incorrect, and she would blush. Because Bonnie does not blush, Carole knows that her assumption that she has a clean forehead is incorrect.

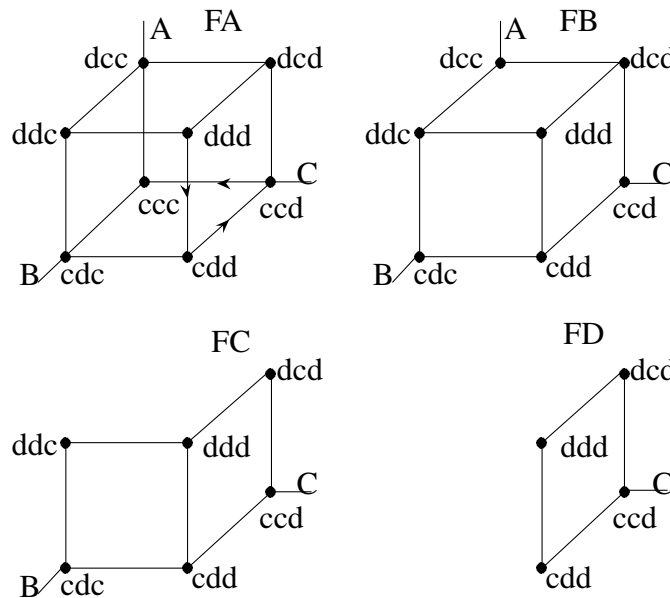


Figure 8.4. The Three Ladies Problem

There is an instructive visual way to approach the problem of the tactful ladies, due to Fagin, Halpern, Moses and Vardi (1995) and illustrated in figure 8.4. Think of each of the ladies owning one of the three axes in this figure, each corner of the cube representing one of the eight states of the world. The endpoints of lines parallel to an axis represent minimal self-

evident events for the lady owning that axis; i.e., the lady in question cannot determine whether her own forehead is dirty.

Because the endpoints of every line segment is a minimal self-evident event for one of the ladies, a node is reachable from another provided there is some path along the lines of the graph, connecting the first to the second. What, for instance, does it mean that ccc is reachable from ddd along the arrows in pane FA of the figure? First, at ddd , Alice believes cdd is possible, at cdd , B believes ccd is possible, and at ccd , C believes that ccc is possible. In other words, at ddd , Alice believes that it is possible that B believes that it is possible that C believes that ccc might be the true state. Indeed, it is easy to see that any sequence of moves around the cube corresponds to some statement of the form x believes it is possible that y believes it is possible that..., and so on. We define an event $E \subseteq \Omega$ as a public event, or common knowledge, if every state $\omega \in E$ is reachable from every other in this manner. Clearly, the only public event is Ω itself.

When the little boy announces b (someone has a dirty forehead), and assuming this statement is taken as truthful, then the three ladies then all know that ccc cannot occur, so we can delete all the paths from some state to ccc . The result is shown in pane FB of the figure. Now, if dcc were the state, Alice would know she has a dirty forehead, and because she apparently does not know this, we can delete the lines terminating in dcc , leading to pane FC in the figure. Now, at ddc or cdc , Bonnie would know she has a dirty forehead, so we can delete the lines connecting to these two nodes. This leaves the nodes depicted in pane FD. Clearly, Carole knows at this event that she has a dirty forehead, but Alice and Bonnie do not.

8.6 The Tactful Ladies and the Commonality of Knowledge

The Three Tactful Ladies problem involves many unstated epistemological assertions going far beyond the common knowledge of rationality involved in the conclusion that Carole knows the state of her forehead. Let us see exactly what they are.

Let x_i be the condition that i has a dirty forehead, and let k_i be the knowledge operator for i , where $i = A, B, C$, standing for Alice, Bonnie, and Carole, respectively. When we write i , we mean any $i = A, B, C$, and when we write i, j , we mean any $i, j = A, B, C$ with $j \neq i$, and when we write i, j, m we mean $i, j, m = A, B, C$ and $i \neq j \neq m \neq i$. Let y_i be the condition that i blushes. The six symbols x_i and y_i represent the possible states

of affairs in a state space Ω . Let E be the event prior to the little boy's exclamation $b = x_A \vee x_B \vee x_C$.

The statement of the problem tells us that $x_i \in E$, and $k_i x_j \in E$; i.e., each lady sees the forehead of the other two ladies, but not her own. The problem also asserts that $k_i x_i \Rightarrow y_i \in E$ (a lady who knows she has a dirty forehead will blush), and $y_i \Rightarrow k_j y_i \in E$. It is easy to check that these conditions are compatible with $\neg k_i x_i \in E$; i.e., no lady knows the state of her own forehead at event E . These conditions also imply that $k_i b \in E$ (each lady knows the little boy's statement is true).

While the problem intends that $k_i x_j \Rightarrow k_i k_m x_j \in E$; i.e., if i knows that j has a dirty forehead, she then knows that m knows this as well, this implication does not follow from any principle of rationality, so we must include it as a new principle. The concept needed is that of a mutually accessible natural occurrence. The mutual accessibility of x_i to j and m may appear to be a weak assumption, but in fact it is the *first time* we have made a substantive assertion that one agent knows that another agent knows something. With this assumption, $k_i k_j b \in E$ follows—each lady knows the others know b holds in E (recall that b is the little boy's statement that ccc is false). To see this, note that $k_i x_j \Rightarrow k_i k_m x_j \Rightarrow k_i k_m b$, which is true for all i and $m \neq i$.

Let E' be the state of knowledge following the exclamation $b = x_A \vee x_B \vee x_C$, which we assume is common knowledge. To prove that in E' one of the ladies (e.g., Carole) blushes, we will assume that y_i is mutually accessible to j, m , and j is a symmetrical reasoner with respect to m concerning event y_i .

The reasoning following the little boy's statement can be summarized as follows. We will show that if Carole assumes $\neg x_C$ at any state in E' , she will arrive at a contradiction. Assuming $\neg x_C$ is true and b is common knowledge, we have $k_C k_B (\neg x_B \Rightarrow k_A \neg x_B \Rightarrow k_A (\neg x_B \wedge \neg x_C \wedge b) \Rightarrow k_A x_A \Rightarrow y_A) \Rightarrow k_C k_B y_A \Rightarrow k_C y_A$, which is false in E' . Thus in E' , $k_C k_B x_B \Rightarrow k_C y_B$, which is not true at any state in E' . Hence x_C is true in E' , and since Carole knows the current state is in E' , $k_C x_C$, so Carole blushes.

8.7 Agreeing to Disagree

In a four page paper buried in the *Annals of Statistics*, Robert Aumann (1976) proved a remarkable theorem. He showed that if two agents have

the same priors concerning an event, and if they update their priors using their private knowledge of the current state ω , and if their posterior probabilities are common knowledge, then these posterior probabilities must be equal. In short, two rational agents with common priors cannot “agree to disagree,” even though the information upon which each bases his updating can be quite different. I will call any theorem with this conclusion an *agreement theorem*. Aumann commented that “We publish this observation with some diffidence, because once one has the appropriate framework, it is mathematically trivial.” (p. 1236). It is valuable to understand this theorem and its generalizations because, as it turns out, the common knowledge conditions for Nash equilibrium are such as to entail an agreement theorem among the agent’s as to how they will play the game.

Suppose Alice and Bob have a common prior p over Ω , where $p(\omega) > 0$ for all $\omega \in \Omega$. Suppose the actual state is ω_α , leading Alice to update the probability of an event E from $p(E)$ to $p_A(E) = p(E|\mathbf{P}_A\omega_\alpha) = a$, and leads Bob to update $p(E)$ to $p_B(E) = p(E|\mathbf{P}_B\omega_\alpha) = b$. Then, if $p_A(E) = a$ and $p_B(E) = b$ are common knowledge, we must have $a = b$. Thus, despite the fact that Alice and Bob may have different information ($\mathbf{P}_A\omega_\alpha \neq \mathbf{P}_B\omega_\alpha$), their posterior probabilities cannot disagree if they are common knowledge.

To see this, suppose the minimal public event containing ω_α is $\mathbf{K}_*^{\omega_\alpha} = \mathbf{P}_A\omega_1 \cup \dots \cup \mathbf{P}_A\omega_k$, where each of the $\mathbf{P}_A\omega_i$ is a minimal self-evident event for Alice. Because the event $p_A(E) = a$ is common knowledge, it is constant on $\mathbf{K}_*^{\omega_\alpha}$, so for any j , $a = p_A(E) = p(E|\mathbf{P}_A\omega_j) = p(E \cap \mathbf{P}_A\omega_j)/p(\mathbf{P}_A\omega_j)$, so $p(E \cap \mathbf{P}_A\omega_j) = ap(\mathbf{P}_A\omega_j)$. Thus,

$$\begin{aligned} p_A(E \cap \mathbf{K}_*^{\omega_\alpha}) &= p(E \cap \cup_i \mathbf{P}_A\omega_i) = p(\cup_i E \cap \mathbf{P}_A\omega_i) \\ &= \sum_i p(E \cap \mathbf{P}_A\omega_i) = a \sum_i p(\mathbf{P}_A\omega_i) = ap(\mathbf{K}_*^{\omega_\alpha}). \end{aligned}$$

However, by similar reasoning, $p_A(E \cap \mathbf{K}_*^{\omega_\alpha}) = bp(\mathbf{K}_*^{\omega_\alpha})$. Hence, $a = b$.

It may seem that this theorem would have limited applicability, because when people disagree, their posterior probabilities are usually private information. But, suppose Alice and Bob are risk-neutral, each has certain financial assets, they agree to trade these assets, and there is a small cost to trading. Let E be the event that the expected value of Alice’s assets is greater than the expected value of Bob’s assets. If they agree to trade then Alice believes E with probability one and Bob believes E with probability zero, and this is indeed common knowledge, because their agreement

indicates that their desire to trade. This is a contradictory situation, which proves that Alice and Bob cannot agree to trade.

Because in the real world, people trade financial assets every day in huge quantities, this proves that either common knowledge of rationality or common priors must be false. In fact, both are probably false. As I argued in §5.11, a rational agent will violate CKR whenever such a violation will increase his expected payoff, a situation that is often the case where the subgame perfect equilibrium has relatively low payoffs for the players (§5.9,5.7). Moreover, there is little reason to believe that the Harsanyi Doctrine (§7.7) holds with respect to stock market prices (Kurz 1997).

We can generalize Aumann's argument as considerably. Let $f(P)$ be a real number for every $P \subseteq \Omega$. We say a f satisfies the *sure thing principle* on Ω if for all $P, Q \subseteq \Omega$ with $P \cap Q = \emptyset$, if $f(P) = f(Q) = a$, then $f(P \cup Q) = a$. For instance, if p is a probability distribution on Ω and E is an event, then the posterior probability $f(X) = p(E|X)$ satisfies the sure thing principle, as does the expected value $f(X) = \mathbf{E}[x|X]$ of a random variable x given $X \subseteq \Omega$. We then have the following *agreement theorem* (Collins 1997):

THEOREM 8.3 *Suppose for each agent $i = 1, \dots, n$, f_i satisfies the sure thing principle on Ω , and suppose it is common knowledge at ω that $f_i = s_i$. Then $f_i(\mathbf{K}_*^\omega) = s_i$ for all i , where \mathbf{K}_*^ω is the cell of the common knowledge partition that contains ω .*

PROOF: To prove the theorem, note that \mathbf{K}_*^ω is the disjoint union of i 's possibility sets $\mathbf{P}_i\omega'$, and $f_i = s_i$ on each of these sets. Hence, by the sure thing principle, $f_i = s_i$ on \mathbf{K}_*^ω . ■

COROLLARY 8.3.1 *Suppose agents $i = 1, \dots, n$ have share a common prior on Ω , indicating an event E has probability $p(E)$. Suppose each agent i now receives private information that the actual state ω is in $\mathbf{P}_i\omega$. Then, if the posterior probabilities $s_i = p(E|\mathbf{P}_i\omega)$ are common knowledge, then $s_1 = \dots = s_n$.*

COROLLARY 8.3.2 *Suppose rational, risk-neutral agents $i = 1, \dots, n$ have the same subjective prior p on Ω , and each has a portfolio of assets X_i , all of which have equal expected value $\mathbf{E}_p(X_1) = \dots = \mathbf{E}_p(X_n)$, and there is a small trading cost $\epsilon > 0$, so no pair of agents desires to trade. In state ω , where agents have posterior expected values $\mathbf{E}_p(X_i|\mathbf{P}_i\omega)$, it cannot be common knowledge that an agent desires to trade.*

Finally, we come to our sought-after relationship between common knowledge and Nash equilibrium:

THEOREM 8.4 *Let \mathcal{G} be an epistemic game with $n > 2$ players, and let $\phi = \phi^1, \dots, \phi^n$ be a set of conjectures. Suppose the players have a common prior p , all players are rational at $\omega \in \Omega$, and it is commonly known at ω that ϕ is the set of conjectures for the game. Then for each $j = 1, \dots, n$, all $i \neq j$ induce the same conjecture $\sigma_j(\omega)$ about j , and $(\sigma_1(\omega), \dots, \sigma_n(\omega))$ form a Nash equilibrium of \mathcal{G} .*

The surprising aspect of this theorem is that if conjectures are common knowledge, they must be independently distributed. This is true, essentially, because it is assumed that a player's prior ϕ_i^ω is independent from his own action $s_i(\omega)$. Thus, when strategies are common knowledge, they can be correlated, but their conditional probabilities given ω must be independently distributed.

PROOF: To prove this theorem, we note that by (8.3), $\phi = \phi^1, \dots, \phi^n$ are common knowledge at ω , and hence $(\sigma_1(\omega), \dots, \sigma_n(\omega))$ are uniquely defined. Because all agents are rational at ω , each $s_i(\omega)$ maximizes $\mathbf{E}[\pi_i(s_i, \phi_i^\omega)]$. It remains only to show that the conjectures imply that agent strategies are uncorrelated. Let $F = \{\omega' | \phi^{\omega'} \text{ is common knowledge}\}$. Because $\omega \in F$ and $p(\mathbf{P}\omega) > 0$, we have $p(F) > 0$. Now let $Q(a) = P([s] | F)$ and $Q(s_i) = P([s_i] | F)$, where in general, we define $[x] = \{\omega \in \Omega | x(\omega) = x\}$ for some variable function $x : \Omega \rightarrow \mathbf{R}$ (thus, $[s] = \{\omega \in \Omega | s(\omega) = s\}$). Note that $Q(a) = P([s] \cap F) / P(F)$. Now let $H_i = [s_i] \cap F$. Because F is commonly known and $[s_i]$ is known to i , H_i is known to i . Hence H_i is the union of minimal i -known events of the form $\mathbf{P}_i \omega'$, and $p([s_i] \cap \mathbf{P}_i \omega') = \phi_i^\omega(s_{-i}) p(\mathbf{P}_i \omega')$. Adding up over all the $\mathbf{P}_i \omega'$ comprising H_i (a disjoint union), we conclude $P([s] \cap F) = P([s_{-i}] \cap H) = \phi_i^\omega(s_{-i}) P(H_i) = \phi_i^\omega(s_{-i}) Q(s_i) P(F)$. Dividing by $P(F)$, we get $Q(a) = \phi_i^\omega(s_{-i}) Q(s_i) = Q(s_{-i}) Q(s_i)$.

It remains to prove that if $Q(a) = Q(s_{-i}) Q(s_i)$ for all $i = 1, \dots, n$, then $Q(a) = Q(s_1) \cdots Q(s_n)$. This is clearly true for $n = 1, 2$. Suppose it is true for $n = 1, 2, \dots, n - 1$. Starting with $Q(a) = Q(s_1) Q(s_{-1})$ where $a = (s_1, \dots, s_n)$, we sum over s_i , getting $Q(s_{-n}) = Q(s_1) Q(s_2, \dots, s_{n-1})$. Similarly, $Q(s_{-i}) = Q(s_i) Q(s_2, \dots, s_{i-1}, s_{i+1}, \dots, s_{n-1})$ for any $i = 1, \dots, n - 1$. By the induction hypothesis, $Q(s_{-n}) = Q(s_1) Q(s_2) \cdots Q(s_{n-1})$, so $Q(a) = Q(s_1) \cdots Q(s_n)$. ■

Theorem 8.4 indicates that common priors and common knowledge of conjectures are the epistemic conditions we need to conclude that rational agents will implement a Nash equilibrium. The question, then, is under what conditions are common priors and common knowledge of conjectures likely to be instantiated in real-world strategic interactions?

8.8 The Demise of Methodological Individualism

There is a tacit understanding among classical game theorists that no information other than the rationality of the agents should be relevant to analyzing how they play a game. This understanding is a form of *methodological individualism*, a doctrine that holds that social behavior consists of the interaction of individuals, so nothing beyond the characteristics of individuals is needed, or even permitted, in modeling social behavior.

The most prominent proponent of methodological individualism was Austrian school economist and philosopher Ludwig von Mises, in his book *Human Action*, first published in 1949. While most of Austrian school economic theory has not stood the test of time, methodological individualism has, if anything, grown in stature among economists, especially since the “rational expectations” revolution in macroeconomic theory (Lucas 1981). “Nobody ventures to deny,” writes von Mises, “that nations, states, municipalities, parties, religious communities, are real factors determining the course of human events.” He continues (p. 42):

Methodological individualism, far from contesting the significance of such collective wholes, considers it as one of its main tasks to describe and to analyze their becoming and their disappearing, their changing structures, and their operation.

von Mises arguments in favor of this principle involve an appeal neither to social theory nor social fact. Rather, he asserts, “a social collective has no existence and reality outside of the individual members’ actions... the way to a cognition of collective wholes is through an analysis of the individuals’ actions.” (p. 42).

This defense, of course, is merely a restatement of the principle. A passing familiarity with levels of explanation in natural science shows that it is not *prima facie* plausible. A computer, for instance, is composed of a myriad of solid state and other electrical and mechanical devices, but the statement that one can successfully model the operation of a computer using only models of the behavior of these underlying parts is just false, even in

principle. Similarly, eukaryotic cells are composed of a myriad of organic chemicals, yet organic chemistry does not supply all the tools for modeling cell dynamics.

We learn from modern complexity theory that there are many levels of physical existence on earth, from the elementary particles to human beings, each level solidly grounded in the interaction of entities of a lower level, yet having emergent properties that are ineluctably associated with the dynamic interaction of its lower level constituents, yet are incapable of being explained on a lower level. The panoramic history of life synthesis of biologists Maynard Smith and Szathmáry (1997) elaborates this theme that every major transition in evolution has taken the form of a higher level of biological organization exhibiting properties that cannot be deduced from its constituent parts. Morowitz (2002) extends the analysis to emergence in physical systems. Indeed, the point should not be mystifying, because there is nothing preventing the most economical model of a phenomenon being the model itself (Chaitin 2004). Adding emergent properties as fundamental entities in the higher level model thus may permit the otherwise impossible: the explanation of complex phenomena.

Epistemic game theory suggests that the conditions ensuring that individuals play a Nash equilibrium are not limited to their *personal* characteristics, but rather include their *common* characteristics, in the form of common priors and common knowledge. We saw (Theorem 7.2) that both individual characteristics and collective understandings, the latter being irreducible to individual characteristics, are needed to explain common knowledge. It is for this reason that methodological individualism is incorrect when applied to the analysis of social life.

Game theory has progressed by accepting no conceptual constructs above the level of the individual actor, as counceled by methodological individualism. Social theory operating at a higher level of aggregation, such as much sociological theory, has produced important insights but has not developed an analytical core on which solid cumulative explanatory progress can be based. The material presented here suggests the fruitfulness of dropping methodological individualist ideology, but carefully articulating the analytical linkages between individually rational behavior and the social institutions that align the beliefs and expectations of individuals, making possible effective social intercourse.

Methodological individualism is inadequate, ultimately, because human nature in general, and human rationality in particular, are products of bio-

logical evolution. The evolutionary dynamic of human groups has produced *social norms* that coordinate the strategic interaction of rational individuals and regulate kinship, family life, the division of labor, property rights, cultural norms, and social conventions. It is a mistake (the error of methodological individualism) to think that social norms can be brought within the purview of game theory by reducing a social institution to the interaction of rational agents.